Recent Development of Linear Scaling Quantum Theories in GAMESS

Cheol Ho Choi

Department of Chemistry, Kyungpook National University, Daegu 702-701, Korea
Received February 28, 2003

Linear scaling quantum theories are reviewed especially focusing on the method adopted in GAMESS. The three key translation equations of the fast multipole method (FMM) are deduced from the general polypolar expansions given earlier by Steinborn and Ruedenberg. Simplifications are introduced for the rotation-based FMM that lead to a very compact FMM formalism. The OPS (optimum parameter searching) procedure, a stable and efficient way of obtaining the optimum set of FMM parameters, is established with complete control over the tolerable error ε. In addition, a new parallel FMM algorithm, requiring virtually no inter-node communication, is suggested which is suitable for the parallel construction of Fock matrices in electronic structure calculations.

Key Words : Linear scaling quantum theory, Fast multipole method, Parallel algorithm, Sperical harmonics, *Ab initio*

Introduction

With the help of ever improving computer hardware and new theories, computational chemistry, with electronic structure theory in particular, is evolving into an essential tool in exploring and understanding the nature of chemical systems. At the same time, as shown in Figure 1, it is quite clear that conventional electronic structure theories may not be able to keep pace with the growth of such demands due to the fact that even in the simplest method, the amount of computations needed increases as a high power of system size. In the case of post-HF theories that include electron correlation effects, the scaling becomes between N^5 to N^{10} making the use of these theories prohibitive. The problem is especially apparent when one wants to apply the conventional *ab initio* theories to the study of nano-size systems.

Since the pioneering work by Almlöf and co-workers,[1] a great deal of studies[2] has been devoted to the development of linear scaling or low-scaling *ab initio* theories in order to overcome the scaling barriers of conventional methods.

In this review, a brief introduction to these new theories shall be given by taking our work as a primary example.

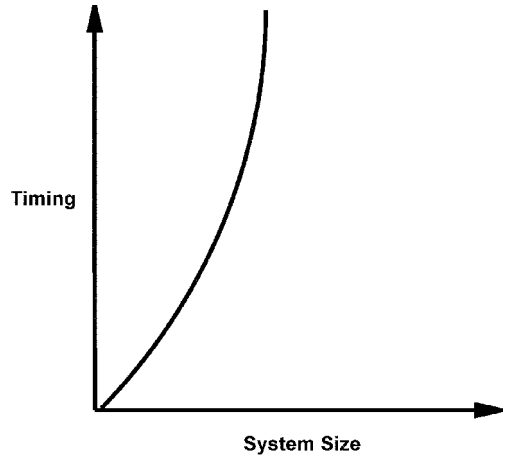

Figure 1. The scaling problem of traditional ab initio theories.

The Basic Idea

The most time consuming part of HF computations is constructing the Fock matrix. That is,

$$F = (h + J - 1/2K) \tag{2.1a}$$

$$h_{ab} = (a|h|b) \tag{2.1b}$$

$$J_{ab} = \sum_c \sum_d D_{cd}(ab|cd) \tag{2.1c}$$

$$K_{ab} = \sum_c \sum_d D_{cd}(ac|bd) \tag{2.1d}$$

where, h is one electron matrix, J and K are Coulomb and exchange matrixes, D_{ab} is the density matrix, and $(ac|bd)$ is two- electron repulsion integrals.

Since the two electron terms formally requires N^4 numbers of the four-center-two-electron repulsion integrals (ERI), the computational demands of Hartree-Fock (HF) method increases with the fourth power of the molecular size. Due to the many screening techniques, the quartic scaling reduces to $N^2(\ln N)$ in the asymptotic region and eventually approaches N^2.[3] However, this is still one of the bottlenecks of *ab initio* computations.

It has been known that one of the most efficient algorithms to construct J and K is to exploit the full permutational symmetry of ERI. Therefore, the conventional procedure to construct J and K begins with the computations of all necessary integrals over primitive gaussian functions at a given shell-quartet. After that, the necessary ERIs are formed and J and K are simultaneously constructed.

J, the Coulomb matrix describing the classical Coulomb potential among electrons, is inversely proportional to r, yielding significant interactions even at long distance. In contrast, K, the exchange matrix is purely a quantum mechanical term used due to the indistinguishability of electrons. Although the distance behavior of K is not well understood, it is generally known that it decreases exponentially in relation to increasing distance, making virtually no contribution from long distance interactions. As a result, the J and K behave

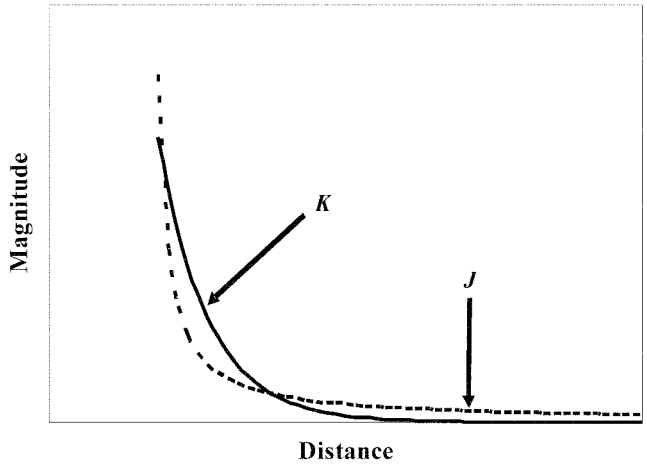

Figure 2. The dotted and the solid lines are the Coulomb and exchange interactions as a function of distance, respectively.

quite differently as a function of distance (see Figure 2).

Consequently, it becomes clear that if one separates the constructions of J and K, then one may be able to derive more specific methods for each of them. In the following sections, the general idea of the linear scaling constructions of J is presented.

Point Charge Approximations of Coulomb Interaction

The Coulomb matrix element can be rewritten as

$$J_{ab} = \sum_c \sum_d D_{cd}(ab|cd) = \left(ab|\sum_c \sum_d D_{cd}cd\right) \tag{3.1a}$$

$$= \left(ab|short\_range\sum_c \sum_d D_{cd}cd\right) + \left(ab|long\_range\sum_c \sum_d D_{cd}cd\right) \tag{3.1b}$$

If we only consider J, then the summation occurs only in *ket* of ERI. Therefore, one can pre-contract the *ket* with the density matrix as in Eq. (3.1a). Suppose the short- and long-range Coulomb interactions can be pre-screened,[4] and then the J can be divided into two parts as written in Eq. (3.1b). The short-range J interaction must be evaluated with the conventional integral code, since any approximation would yield uncontrollable error. However, as Figure 3 illustrates, the long-range J interaction can be approximated by point charge interaction as

$$(ab|D_{cd}cd) \cong \rho_{ab}\rho_{cd}/r_{12} \tag{3.2}$$

This approximation is valid only if the distribution range of the product of the Gaussian functions is much smaller than the distance between the distributions. Point charge approximation by Eq. (3.2) alone cannot yield linear scaling method, since the Coulomb potential energies among N point charges still requires N^2 amount of computations due to the pair-wise interactions. Therefore, one has to introduce a fast method to reduce the quadratic scaling of Coulomb potential energy evaluations among point charges.

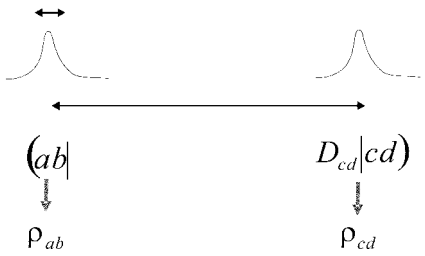

Figure 3. Illustration of point charge approximation of ERI. The product of Gaussian functions can be approximated as a point charge.

Multipole Expansions of Point Charges

One of the fast method to compute Coulomb interactions among point charges is the fast multipole method (FMM),[5] which has the potential to reduce the $O(N^2)$ work of calculating a pairwise potential to $O(N)$ for N particles.

Many quantum chemists have recognized the potential use of FMM in the quantum mechanical computations yielding many encouraging results, such as QCTC (Quantum Chemical Tree Code),[6] CFMM (Continuous Fast Multipole Method)[7] and GvFMM (Gaussian very Fast Multipole Method).[8] White and Head-Gordon (WH) have contributed many significant improvements,[9] such as the rotation based FMM and the fractional tier method. These authors demonstrated that linear scaling with respect to the number of particles can be achieved, as long as edge effects can be minimized. However, current methods do not have direct explicit control over the error in FMM calculations.

In order to take advantage of FMM, the point charge interaction needs to be expanded with multipoles such as,

$$(\rho_{ab}\rho_{cd}/r_{12}) = \rho_{ab}\rho_{cd}\left|\vec{r}_> - \vec{r}_<\right|^{-1} = \sum_{l=0}^{\infty}\sum_{m=(-l)}^{l} \hat{Y}_l^m(\vec{r}_<)\hat{Z}_l^m(\vec{r}_>) \tag{4.1}$$

where $\hat{Y}_l^m$ and $\hat{Z}_l^m$ are the *regular* and the *irregular* spherical harmonics. Consequently the long range Coulomb interactions become

$$\left(ab|long\_range\sum_c \sum_d D_{cd}cd\right) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \hat{Y}_l^m(\vec{r}_<)\sum_c \sum_d \hat{Z}_l^m(\vec{r}_>) \tag{4.2}$$

Within the double precision, the usual value of the largest l is 20. Therefore the length of the first two summations of Eq. (4.2) is fixed. However, the last two summations of Eq. (4.2) run over basis sets which depend on the system-size. With the help of FMM, the last two summations can be replaced with a fast algorithm yielding linear computations of long range Coulomb interactions. The amount of long range interactions compared to the short-range interactions increases with the system size. Therefore, this method is especially useful for large systems.

Deriving an Efficient FMM Algorithm

The general description of FMM would not be given here, since it is beyond the scope of this current review. Readers are encouraged to refer to Greengard's thesis.[5] Instead, we shall describe the derivations of the necessary equations of FMM. The algorithm requires three different types of the multipole expansion translations.[10] It is not well recognized that the three key translational equations of FMM can be derived from the two general polypolar expansions.[11]

$$\hat{Y}_L^M(\vec{r}_1+\vec{r}_2+\cdots+\vec{r}_n) = \sum_{l_1 m_1}\sum_{l_2 m_2}\sum_{l_3 m_3}\cdots\sum_{l_n m_n}\hat{Y}_{l_1}^{m_1}(\vec{r}_1)\hat{Y}_{l_2}^{m_2}(\vec{r}_2)\cdots\hat{Y}_{l_n}^{m_n}(\vec{r}_n) \quad (5.1a)$$

$$\hat{Z}_L^M(\vec{r}_1+\vec{r}_2+\cdots+\vec{r}_{n-1}) = \sum_{l_1 m_1}\sum_{l_2 m_2}\sum_{l_3 m_3}\cdots\sum_{l_n m_n}(-1)^{l_1-m_1-l_2+m_2+\cdots-l_n m_n} \cdot \hat{Y}_{l_1}^{-m_1}(\vec{r}_1)\hat{Y}_{l_2}^{-m_2}(\vec{r}_2)\cdots\hat{Y}_{l_n}^{-m_n}(\vec{r}_n)\hat{Z}_{L+l_1-l_2+\cdots+l_n}^{M+m_1+m_2+\ldots+m_n}(\vec{r}_{n-1}) \quad (5.1b)$$

Translations of solid harmonics take very simple explicit forms when the z-axis is parallel to the translation vector so that the latter has spherical coordinates $\theta=\phi=0$ in the coordinate system of the harmonics. White and Head-Gordon[7] therefore suggest that it is efficient first to rotate the spherical harmonics such that their z-axes are parallel to the translation axis. The translation is then performed along the z-axis, after which the solid harmonics are rotated back to the original coordinate system. In such a context, the efficient *rotation* of solid harmonics is therefore also a very relevant consideration. Now, a rotation of the coordinate system induces a unitary transformation among the *surface* harmonics, namely

$$Y_L^M(\vec{r}') = \sum_{k=-L}^{L} D_{kM}^L Y_L^M(\vec{r}) \quad (5.2)$$

where the D_{kM}^L are the Wigner rotation matrices that transform the spherical harmonics defined with respect to one coordinate system ($\vec{r}$) into the spherical harmonics defined with respect to the rotated coordinate system ($\vec{r}'$). An efficient and numerically stable recurrence procedure for the rapid evaluation of the Wigner $\boldsymbol{D}^L$ matrix directly from the elements of the coordinate rotation matrix $\boldsymbol{R}$ has been derived.[12]

$$D_{m,m'}^l = a_{m,m'}^l D_{00} D_{m,m'}^{l-1} + b_{m,m'}^l D_{10} D_{m-1,m'}^{l-1} + b_{-m,m'}^l D_{-10} D_{m-1,m'}^{l-1} \quad (5.3a)$$

$$D_{m,m'}^l = c_{m,-m'}^l D_{0,-1} D_{m,m'-1}^{l-1} + d_{m,-m'}^l D_{1,-1} D_{m-1,m'+1}^{l-1} + d_{-m,-m'}^l D_{-1,-1} D_{m+1,m'+1}^{l-1} \quad (5.3b)$$

$$D_{m,m'}^l = c_{m,m'}^l D_{0,1} D_{m,m'-1}^{l-1} + d_{m,m'}^l D_{11} D_{m-1,m'-1}^{l-1} + d_{-m,m'}^l D_{-1,1} D_{m-1,m'-1}^{l-1} \quad (5.3c)$$

On the basis of Eq. (5.1) and considering the rotation based FMM, the most compact translational formulas along a particular direction were derived.[13]

$$\tilde{Y}_L^M(\vec{r}_1+\vec{r}_2) = \sum_{j=|M|}^{L}\frac{|\vec{r}_1|^{L-j}}{(L-j)!}\left(\frac{(L+M)!(L-M)!}{(j+M)!(j-M)!}\right)^{1/2}\tilde{Y}_j^M(\vec{r}_2) \quad (5.4a)$$

$$\tilde{Z}_L^M(\vec{r}_< - \vec{r}_>) = \frac{(-1)^{L-M}}{[(L-M)!(L+M)!]^{1/2}} \sum_{l=|M|}^{\infty}\frac{(L+l)!}{|\vec{r}_>|[(l-M)!(l+M)!]^{1/2}}\tilde{Y}_l^M(\vec{r}_<)^* \quad (5.4b)$$

$$\tilde{Z}_L^M(\vec{r}_> - \vec{r}_<) = \frac{1}{[(L-M)!(L+M)!]^{1/2}} \sum_{l=L}^{\infty}\frac{|\vec{r}_<|^{l-L}[(l-M)!(l+M)!]^{1/2}}{(l-L)!}\tilde{Z}_l^M(\vec{r}_>)^* \quad (5.4c)$$

The equations (5.3) and (5.4) comprise the most efficient and numerically stable formulas for FMM.

The Optimum Parameter Searching (OPS) Procedure

Choosing the optimum set of FMM parameters is not an easy task, since they are inter-related due to the complexity of FMM. White and Head-Gordon[7] have achieved a significant performance improvement using their fractional tier method. Their basic idea is that one can balance the near-field and far-field computational effort by minimizing the variation in the number of particles per lowest level box relative to the optimal value. However, the best choice for the optimum number of particles is not obvious, since one does not have control over numerical accuracy.

In order to derive an OPS procedure, the computational overhead of FMM was assessed. The two major overheads are

$$\text{direct interaction overhead, } T_d = \frac{(2\cdot ws+1)^3 M^2}{2\cdot 8^{N_s}} t_d \quad (6.1)$$

and

$$\text{direct transfer overhead, } T_t = \sum_{n=1}^{N_s} 8^n(2T_{t1}+T_{t2}) = \sum_{n=1}^{N_s} 8^n((2(2ws+1))^3-(2ws+1)^3+2) \cdot (2(2l_{max}+1)^3+(l_{max}+1)^3)t_t \quad (6.2)$$

With these assessments, the optimum number of boxes in

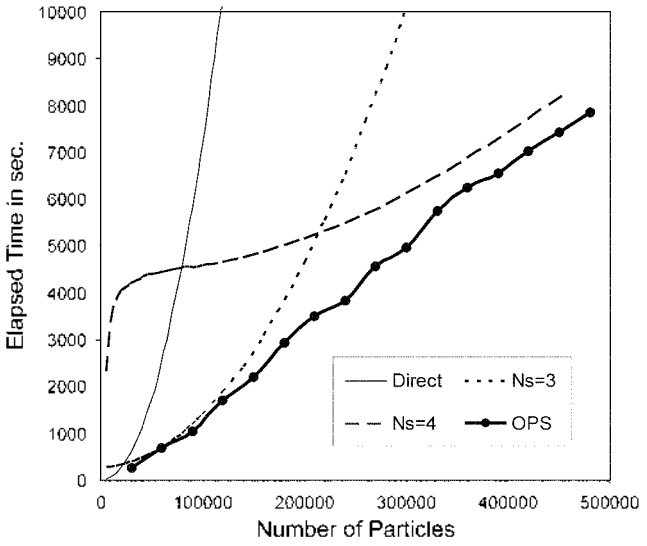

Figure 4. Test runs of conventional and FMM methods. Thin solid line represents the $O(N^2)$ scaling of normal method. The dotted lines represent the FMM results without OPS technique showing the importance of OPS. Ns is the level of subdivision which has significant impact on the performance FMM computations. The thick solid line represents the current method with OPS technique showing the linear scaling performance.

the highest subdivision level, α_0, is determined. That is,

$$\alpha_0 = M\left(\frac{7}{2\cdot 8}\frac{(2ws+1)^3}{((2(2ws+1))^3-(2ws+1)^3+2)}\frac{t_d}{t^f}\cdot(2(2l_{max}+1)^3+(l_{max}+1)^3)\right)^{1/2} \quad (6.3)$$

On the basis of these equations, iterative OPS procedure was derived. In order to illustrate the usefulness of the OPS procedure, test calculations have been performed as a function of the number of randomly distributed unit charges. Plots of time against the number of particles are presented in Figure 4. It is seen that OPS performs always better than other choices of parameters.

A new Parallel FMM Algorithm

A common feature of hierarchical multipole methods is that the particles are recursively subdivided into a hierarchy of boxes, or cells, based on their spatial positions. This hierarchy establishes a tree-like data structure with a complete representation of the particles at each tier, or level, of subdivision. The algorithm begins by dividing a box containing all distributions (parent, root) into 8 equal boxes (children). Recursive sub-division of this type leads to an "oct-tree" data-structure. Each tier of the tree represents the entire set of particles, but involves boxes of increasing spatial resolution formed by sub-division of their parents.

Earlier parallel implementations[14] relied on dividing the boxes of the tree among available processors, a natural way

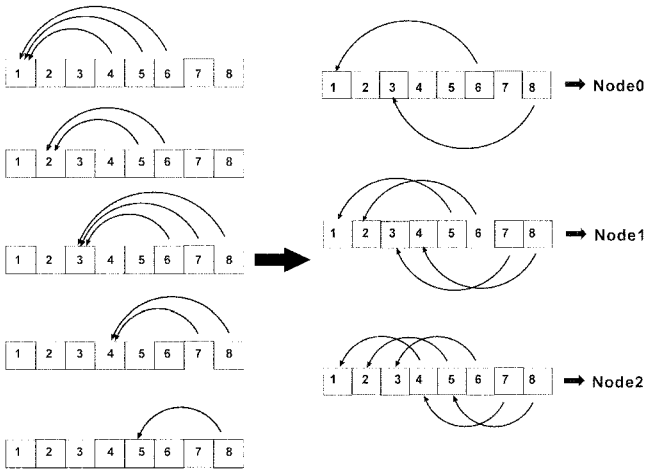

Figure 5. A new parallel FMM algorithm which distributes the types of translations.

to tackle the problem. The main drawback of this approach is that because the data on different compute nodes are not independent of each other, there are significant inter-node communications. The amount of communication increases almost exponentially with the number of processors. This has a serious impact on the scalability of the algorithm and limits the utility of parallel runs. Furthermore, this method can suffer from load-balancing problems if, as is likely, the particles are not uniformly distributed among the boxes. The strategy used in the current algorithm is to distribute the work rather than the data as shown in Figure 5. Left-hand side of Figure 5 shows the conventional sequence of multipole computations. The computation starts from the box 1 and continues to next box until no more necessary computations left. However, our algorithm categorizes the types of computations. Node 0 computes the 1-6 interaction, node 1 computes 1-5 interactions and so on. By distributing the same type of translations, unnecessary computations and the inter-node communications can be removed. Randomly distributed unit charges were used to test the parallel

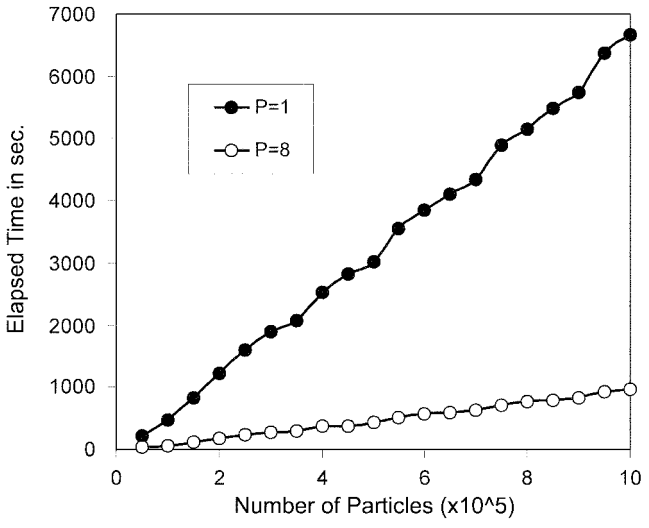

Figure 6. Parallel performance with 1 and 8 processors.

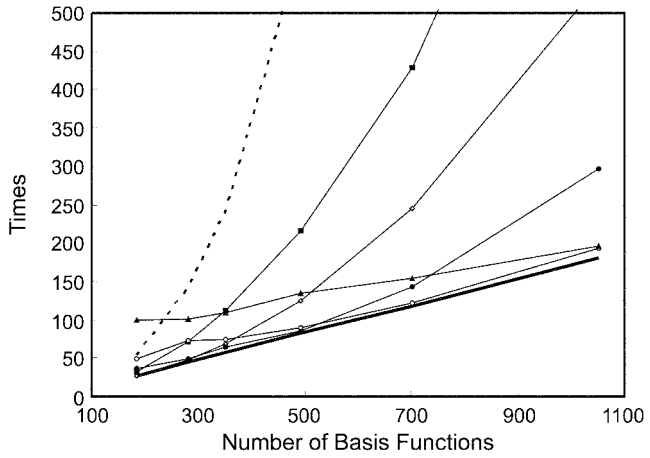

Figure 7. QFMM-HF/STO-3G calculations on the series of oligomers of polyethylene. The dotted line represents the result of conventional HF theory. The solid thick line represents QFMM-HF results with OPS technique. The thin solid lines represent QFMM-HF results at Fixed Ns values which are not determined by OPS OPS technique. The Ns value ranges from 3 to 7.

performance of the new algorithm. Timings with 1 and 8 processors (P) are presented in Figure 6. The plot clearly shows linear scaling with respect to number of particles and parallelization does not degrade the linearity. The actual speedup on 8 processors is 6.9.

The Real Test

On the basis of previously discussed improvements, we have developed Quantum version of FMM (QFMM). To illustrate the performance of QFMM, a series of test runs are presented in Figure 7. The QFMM-HF single point energy calculations were done on a series of oligomers of polyethylene with STO-3G basis set. The dotted line represents the performance of conventional methods, while the solid thick line represents the results of QFMM-HF with the OPS showing the near linear scaling with the basis set size. Our OPS technique not only optimizes performance but also guarantee accuracy. Therefore, the accuracy of the QFMM results is the same as the conventional method. The thin solid lines represent the results of QFMM-HF without OPS indicating the importance of OPS to achieve true linear scaling.

Conclusions

In this short review, basic features of linear scaling methods are presented focusing on the work done by the current author. It has been shown that the three key translation equations can be readily deduced from the generalized equations. Basic regular and irregular harmonics are redefined yielding a compact formalism. The current formalism is advantageous, since $(2L + 1)$ terms are required to perform a rotation, while less than L terms are needed to perform a translation along a specialized axis. The rotation based FMM can be further simplified yielding a very compact FMM formalism.

A new parallel FMM algorithm is introduced that does not require inter-node communication. This new method is suitable for the parallel construction of the Fock matrix of quantum calculations. Instead of assigning divided space to compute-nodes, unique translation types are assigned to compute-nodes for the parallelization of the FMM. The former approach suffers from rapidly growing inter-node communications as a function of space subdivision, and from load-balancing problems. In contrast, the implementation described in the current work does not suffer from these problems, so it is expected to provide a more robust performance. In addition, our implementation can in principle work with any number of processors, limited only by the scaling as the number of processors grows.

A stable and efficient method for determining the optimum set of FMM parameters (OPS) via a user specified acceptable error has been established using a self-consistent process to ensure both accuracy and optimal performance. Since the computational overhead depends most strongly on the number of boxes, the iteration converges very rapidly. The new self-consistent procedure achieves linear scaling with respect to the number of particles with complete control over the actual error.

Potential applications of linear scaling quantum methods are numerous. Carbon nanotubes, fullerene derivatives, drug design,[15] conducting polymers, molecular recognition of proteins, quasi-crystals, and dendrimers are just a few examples of applications that are ripe for exploration.

Acknowledgements. This work was supported by Korea Research Foundation Grant (KRF-2002-070-C00048).

References

1. Panas, I.; Almlöf, J.; Feyereisen, M. W. *Int. J. Quantum Chem.* **1991**, *40*, 797.
2. Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 4.
3. Haser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104.
4. There are two types of error to consider. They are FMM error and the Gaussian distribution range error. The FMM error arises due to the incomplete summations of multipole expansions, that is, $\varepsilon = |\Phi_{FMM} - \Phi_{exact} = \left| \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} \frac{(l-m)!}{(l+m)!} \frac{q_a q_b r_<^l}{r_>^{l+1}} P_l^m(\cos\theta_<) P_l^m(\cos\theta_>) e^{im(\phi_> - \phi_<)} - \frac{q_a q_b}{|\vec{r}_> - \vec{r}_<|} \right|$. The Gaussian distribution range is due to the different exponent of Gaussian distributions.
5. Greengard, L. *The Rapid Evaluation of Potential Fields in Particle Systems*; MIT: Cambridge, 1987.
6. (a) Challacombe, M.; Schwegler, E.; Almlof, J. *J. Chem. Phys.* **1996**, *104*(12), 4685. (b) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*(7), 2726. (c) Challacombe, M. *J. Chem. Phys.* **2000**, *113*(22), 10037.
7. (a) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.*, **1994**, *230*(1-2), 8. (b) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*(3,4), 268. (c) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*(5), 1663.
8. (a) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science (Washington, D.C)* **1996**, *271*(5245), 51. (b) Burant, J. C.; Strain,

M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *248*(1,2), 43. (c) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *258*(1,2), 45.
9. (a) White, C.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593. (b) *ibid. Chem. Phys. Lett.* **1996**, *257*, 647. (c) *ibid. J. Chem. Phys.* **1996**, *105*, 5061.
10. It is in fact, the translations of the expansion center.
11. Steinborn, E. O.; Ruedenberg, K. *Adv. Quantum Chem.* **1973**, *7*, 1.
12. Choi, C. H.; Ivanic, J.; Gordon, M. S.; Ruedenberg, K. *J. Chem. Phys.* **1999**, *111*, 8825.
13. Choi, C. H.; Ruedenberg, K.; Gordon, M. S. *J. Comput. Chem.* **2001**, *22*, 1484.
14. (a) Schmidt, K. E.; Lee, M. A. *J. Stat. Phys.* **1991**, *63*, 1223. (b) Greengard, L.; Gropp, W. D. *Computers Math. Applic.* **1990**, *20*, 63.
15. Ra, C. S.; Park, G. *Bull. Korean Chem. Soc.* **2002**, *23*, 1199.